Project

Design and Implement your Machine Learning Algorithm on the following Dataset.

Projects (Deadline: November 30, 2023):

Complete Dataset Download

C1. Classification

Classification: Classification is to identify which category a new observation belongs to, on the basis of a training dataset. There are five datasets. For each dataset, we provide the training dataset, training label, and test dataset. Please use the training dataset and training label to build your classifier and predict the test label. A class label is represented by an integer. For example, in the 1st dataset, there are 4 classes where 1 represents the 1st class, 2 represents the 2nd class, etc. Note that, there exist some missing values in some of the dataset (a missing entry is filled by 1.00000000000000e+99), please fill the missing values before perform your classification algorithm.

TrainData 1 contains 3312 features with 150 samples. Testdata1 contains 3312 features with 53 samples. There are 5 classes in this dataset.

TrainData 2 contains 9182 features with 100 samples. Testdata2 contains 9182 features with 74 samples. There are 11 classes in this dataset.

TrainData 3 contains 13 features with 6300 samples. Testdata3 contains 13 features with 2693 samples. There are 9 classes in this dataset.

TrainData 4 contains 112 features with 2547 samples. Testdata4 contains 112 features with 1092 samples. There are 9 classes in this

dataset.

TrainData 5 contains 11 features with 1119 samples. Testdata5 contains 11 features with 480 samples. There are 6 classes in this dataset.

TrainData 6 contains 142 features with 612 samples. Testdata6 contains 142 features with 262 samples. This is not a classification problem. You are asked to predict the real value. (Graduate Students Only)

TrainData 1	TrainLabel 1	TestData 1
TrainData 2	TrainLabel 2	TestData 2
TrainData 3	TrainLabel 3	TestData 3
TrainData 4	TrainLabel 4	TestData 4
TrainData 5	TrainLabel 5	TestData 5
TrainData 6	TrainLabel 6	TestData 6

Sample Data:

Training data:

1.1	2.1	2.1	5.2
2.1	2.4	2.4	2.1
3.1	1.5	2.6	1.5

Training label

1
1
2

Test data

3.1	2.2	1.5	2.5
2.1	2.1	2.1	2.6

Please use the training data and training label to predict the test label. For example, if your prediction for the test sample is 1, 2. That is, the first sample in the test dataset (first row) is predicted as 1 and second as 2. Then please return me the test result of each dataset as an individual files.

1
2

2. Missing Value Estimation

Gene expression data often contain missing expression values and it is very important to estimate those missing value as accurate as possible. The first task of the course project is to estimate missing value in the Microarray Data.

Dataset 1 contains 242 genes with 14 samples.

Dataset 2 contains 758 genes with 50 samples.

Dataset 3 contains 273 viruses with 79 samples. There are only 3815 observed values. (Bonus Questions for Undergraduate)

1	1.00000000000000e+99	1.00000000000000e+99
1	1	1
2	2	2

Table 1

Note that the missing entry is filled by 1.00000000000000e+99. For example, in the Table 1, the second and third entries in the first row are missing values. There are 4% missing values in the Dataset 1 and 10% missing values in the Dataset 2. Please fill those missing entries with estimated values and return the complete dataset to me.

3. Multi-label Classification: Multi-label classification is a variant of the classification problem where multiple target labels must be assigned to each sample (Graduate Students Only)

MultLabelTrainData contains 103 features with 500 samples. MultLabelTestData contains 103 features with 100 samples. The label file for the train data can be download at MultLabelTrainLabel

In the following dataset, there are totally 14 target labels. The samples in the training dataset are assigned with more than one target label. For

example, in the first sample MultLabelTrainLabel, the label assignment for the first sample is 7, 8, 12 and 13. Those positions are marked with 1.

0	0	0	0	0	0	1	1	0

Please predict the labels for the test samples. The output file format of Testing Label should be consistent with MultLabelTrainLabel. For example, if there are 3 test samples where the predicting labels for the first sample is that it has label of 2, 3, the predicting label for the second sample is 12, 14, and the predicting labels for the third one is 2, 5. The output is as follows:

0	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0	0

Project Submission (One zip file):

1. Project Report: Please submit a project report which describe your methods for both classifications and missing value estimation (At most 3 pages).

2. The source code and a readme file on how to run the code.

3. Prediction result: Please submit 4(or 5) individual files. Each one is the results return by your program. The file name should start with your last name. For example, CaiClassification1.txt,CaiClassification2.txt, CaiClassification3.txt, CaiMissingResult1.txt, CaiMissingResult2.txt, etc